Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages

Oliver Ferschke[‡], Iryna Gurevych[†‡] and Yevgen Chebotar[‡]

† Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information

‡ Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science
Technische Universität Darmstadt

http://www.ukp.tu-darmstadt.de

Abstract

In this paper, we propose an annotation schema for the discourse analysis of Wikipedia Talk pages aimed at the coordination efforts for article improvement. We apply the annotation schema to a corpus of 100 Talk pages from the Simple English Wikipedia and make the resulting dataset freely available for download[1]. Furthermore, we perform automatic dialog act classification on Wikipedia discussions and achieve an average F_1-score of 0.82 with our classification pipeline.

1 Introduction

Over the past decade, the paradigm of information sharing in the web has shifted towards participatory and collaborative content production. Texts are no longer exclusively prepared by individuals and then shared with the community. They are increasingly created collaboratively by multiple authors and iteratively revised by the community.

When researchers first conducted surveys on professional writers in the 1980s, they found that the collaborative writing process differs considerably from the way individual writing is done (Posner and Baecker, 1992). In joint writing, the writers have to externalize processes that are otherwise not made explicit, like the planning and the organization of the text. The authors have to communicate *how* the text should be written and *what* exactly it should contain.

Today, many tools are available that support collaborative writing. A tool that has particularly taken hold is the *Wiki*, a web-based, asynchronous co-authoring tool. A unique characteristic of Wikis is the documentation of the edit history which keeps track of every change that is made to a Wiki page. With this information, it is possible to reconstruct the writing process from the beginning to the end. Additionally, many Wikis offer their users a communication platform, the Talk pages, where they can discuss the ongoing writing process with other users.

The most prominent example for a successful, large-scale Wiki is *Wikipedia*, a collaboratively created online encyclopedia, which has grown considerably since its launch in 2001, and contains a total of almost 20 million articles in 282 languages and dialects, as of Sept. 2011. As there is no editorial body that manages Wikipedia top-down, it is an open question how the huge online community around Wikipedia regulates and enforces standards of behavior and article quality. The user discussions on the article Talk pages might shed light on this issue and give an insight into the otherwise hidden processes of collaboration that, until now, could only be analyzed via interviews or group observations in experimental settings.

The main goal of the present paper is to analyze the content of the discussion pages of the Simple English Wikipedia with respect to the dialog acts aimed at the coordination efforts for article improvement. Dialog acts, according to the classic speech act theory (Austin, 1962; Searle, 1969), represent the meaning of an utterance at the level of illocutionary force, i.e. a dialog act label concisely characterizes the intention and the role of a contribution in a dialog. We chose the Simple English Wikipedia for our initial analysis, because we are able to obtain more representative results

[1] http://www.ukp.tu-darmstadt.de/data/wikidiscourse

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 777–786,
Avignon, France, April 23 - 27 2012.

by covering almost 15% of all relevant Talk pages, as opposed to the much smaller fraction we could achieve for the English Wikipedia. The long-term goal of this work is to identify relations between contributions on the Talk pages and particular article edits. We plan to analyze the relation between article discussions and article content and identify the edits in the article revision history that react to the problems discussed on the Talk page. In combination with article quality assessment (Yaari et al., 2011), this opens up the possibility to identify successful patterns of collaboration which increase the article quality. Furthermore, our work will enable practical applications. By augmenting Wikipedia articles with the information derived from automatically labeled discussions, article readers can be made aware of particular problems that are being discussed on the Talk page "behind the article".

Our primary contributions in this paper are: (1) an annotation schema for dialog acts reflecting the efforts for coordinating the article improvement; (2) the Simple English Wikipedia Discussion (SEWD) corpus, consisting of 100 segmented and annotated Talk pages which we make freely available for download; and (3) a dialog act classification pipeline that incorporates several state of the art machine learning algorithms and feature selection techniques and achieves an average F_1-score of .82 on our corpus.

2 Related Work

The analysis of speech and dialog acts has its roots in the linguistic field of pragmatics. In 1962, John Austin shifted the focus from the mere declarative use of language as a means for making factual statements towards its non-declarative use as a tool for performing actions. The speech act theory was further systematized by Searle (1969), whose classification of illocutionary acts (Searle, 1976) is still used as a starting point for creating dialog act classification schemata for natural language processing.

A well known, domain- and task-independent annotation schema is DAMSL (Core and Allen, 1997). It was created as the standard annotation schema for dialog tagging on the utterance level by the Discourse Resource Initiative. It uses a four-dimensional tagset that allows arbitrary label combinations for each utterance. Jurafsky et al. (1997) augmented the DAMSL schema to fit the peculiarities of the Switchboard corpus. The resulting SWDB-DAMSL schema contained more than 220 distinct labels which have been clustered to 42 coarse grained labels. Both schemata have often been adapted for special purpose annotation tasks.

With the rise of the social web, the amount of research analyzing user generated discourse substantially increased. In addition to analyzing web forums (Kim et al., 2010a), chats (Carpenter and Fujioka, 2011) and emails (Cohen et al., 2004), Wikipedia Talk pages have recently moved into the center of attention of the research community.

Viégas et al. (2007) manually annotate 25 Wikipedia article discussion pages with a set of 11 labels in order to analyze how Talk pages are used for planning the work on articles and resolving disputes among the editors. Schneider et al. (2011) extend this schema and manually annotate 100 Talk pages with 15 labels. They confirm the findings of Viégas et al. that coordination requests occur most frequently in the discussions.

Bender et al. (2011) describe a corpus of 47 Talk pages which have been annotated for authority claims and alignment moves. With this corpus, the authors analyze how the participants in Wikipedia discussions establish their credibility and how they express agreement and disagreement towards other participants or topics.

From a different perspective, Stvilia et al. (2008) analyze 60 discussion pages in regard to how information quality (IQ) in Wikipedia articles is assessed on the Talk pages and which types of IQ problems are identified by the community. They describe a Wikipedia IQ assessment model and map it to established frameworks. Furthermore, they provide a list of IQ problems along with related causal factors and necessary actions which has also inspired the design of our annotation schema.

Finally, Laniado et al. (2011) examine Wikipedia discussion networks in order to capture structural patterns of interaction. They extract the thread structure from all Talk pages in the English Wikipedia and create tree structures of the discussion. The analysis of the graphs reveals patterns that are unique to Wikipedia discussions and might be used as a means to characterize different types of Talk pages.

To the best of our knowledge, there is no work yet that uses machine learning to automati-

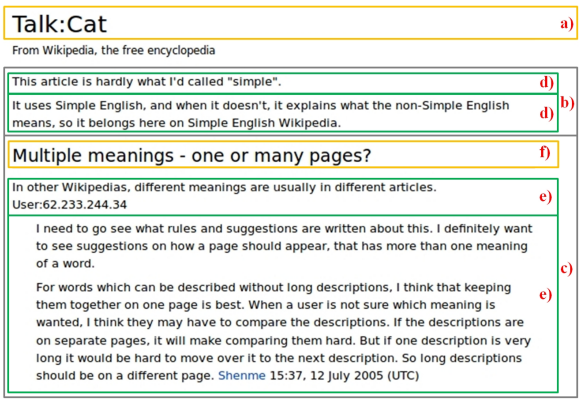

Figure 1: Structure of a Talk page: *a)* Talk page title, *b)* untitled discussion topic, *c)* titled discussion topic, *d)* unsigned turns, *e)* signed turns, *f)* topic title

cally classify user contributions in Wikipedia Talk pages. Furthermore, there is no corpus available that reflects the efforts of article improvement in Wikipedia discussions. This is the subject of our work.

3 Annotation Schema

The main purpose of Wikipedia Talk pages is the coordination of the editing process with the goal of improving and sustaining the quality of the respective article. The criteria for article quality in Wikipedia are loosely defined in the guidelines for *"good articles"*[2] and *"very good articles"*[3]. According to these guidelines, distinguished articles must be *well-written in simple English*, *comprehensive*, *neutral*, *stable*, *accurate*, *verifiable* and follow the Wikipedia *style guidelines*[4]. These criteria are the main points of reference in the discussions on the Talk pages.

Discourse analysis, as it is performed in this paper, can be carried out on various levels, depending on what is regarded as the smallest unit of the discourse. In this work, we focus on turns, not on individual utterances, as we are interested in a coarse-grained analysis of the discourse-structure as a first step towards a finer-grained discourse analysis. We define a *turn* (or contribution) as the body of text that is added by an individual contributor in one or more revisions to a single discussion topic until another contributor edits the page. Furthermore, a *topic* (or discussion) is the body of turns that revolve around a single matter. They are usually headed by a topic title. Finally, the *thread structure* designates the sequence of turns and their indentation levels on the Talk page. A structural overview of a Talk page and its constituents can be seen in Figure 1.

We composed an annotation schema that reflects the coordination efforts for article improvement. Therefore, we manually analyzed a set of thirty Talk pages from the Simple English Wikipedia to identify the types of article deficiencies that are discussed and the way article improvement is coordinated. We furthermore incorporated the findings from an information-scientific analysis of information quality in Wikipedia (Stvilia et al., 2008), which identifies twelve types of quality problems, like e.g. *Accuracy*, *Completeness* or *Relevance*. Our resulting tagset consists of 17 labels (cf. Table 1) which can be subdivided into four higher level categories:

Article Criticism Denote comments that identify deficiencies in the article. The criticism can refer to the article as a whole or to individual parts of the article.

Explicit Performative Announce, report or suggest editing activities.

Information Content Describe the direction of the communication. A contribution can be used to communicate new information to others (`IP`), to request information (`IS`), or to suggest changes to established facts (`IC`). The `IP` label applies to most of the contributions as most comments provide a certain amount of new information.

Interpersonal Describe the attitude that is expressed towards other participants in the discussion and/or their comments.

Since a single turn may consist of several utterances, it can consequently comprise multiple dialog acts. Therefore, we designed the annotation study as a multi-label classification task, i.e. the annotators can assign one or more labels to each annotation unit. Each label is chosen independently. Table 1 shows the labels, their respective definitions and an example from our corpus.

4 Corpus Creation and Analysis

The SEWD corpus consists of 100 annotated Talk pages extracted from a snapshot of the Simple En-

[2] `http://simple.wikipedia.org/wiki/WP:RGA`
[3] `http://simple.wikipedia.org/wiki/WP:RVGA`
[4] `http://simple.wikipedia.org/wiki/WP:STYLE`

Label	Description	Example
Article Criticism		
CM	Content incomplete or lacking detail	*It should be added (1) that voters may skip preferences, but (2) that skipping preferences has no impact on the result of the elections.*
CW	Lack of accuracy or correctness	*Kris Kringle is NOT a Germanic god, but an English mispronunciation of Christkind, a German word that means "the baby Jesus".*
CU	Unsuitable or unnecessary content	*The references should be removed. The reason: The references are too complicated for the typical reader of simple Wikipedia.*
CS	Structural problems	*Also use sectioning, and interlinking*
CL	Deficiencies in language or style	*This section needs to be simplified further; there are a lot of words that are too complex for this wiki.*
COBJ	Objectivity issues	*This article seems to take a clear pro-Christian, anti-commercial view.*
CO	Other kind of criticism	*I have started an article on Google. It needs improvement though.*
Explicit Performative		
PSR	Explicit suggestion, recommendation or request	*This section needs to be simplified further*
PREF	Explicit reference or pointer	*Got it. The URL is* `http://www.dmbeatles.com/history.php?year=1968`
PFC	Commitment to an action in the future	*Okay, I forgot to add that, I'll do so later tonight.*
PPC	Report of a performed action	*I took and hopefully simplified the "[[en:Prehistoric music—Prehistoric music]]" article from EnWP*
Information Content		
IP	Information providing	*"Depression" is the most basic term there is.*
IS	Information seeking	*So what kind of theory would you use for your music composing?*
IC	Information correcting	*In linguistics and generally speaking, when Talking about the lexicon in a language, words are usually categorized as 'nouns', 'verbs', 'adjectives' and so on. The term 'doing word' does not exist.*
Interpersonal		
ATT+	Positive attitude towards other contributor or acceptance	*Thank you.*
ATTP	Partial acceptance or partial rejection	*Okay, I can understand that, but some citations are going to have to be included for [[WP:V]].*
ATT-	Negative attitude towards other contributor or rejection	*Now what? You think you know so much about everything, and you are not even helping?!*

Table 1: Annotation schema for the dialog act classification in Wikipedia discussion pages with examples from the SEWD Corpus. Some examples have been shortened to fit the table.

glish Wikipedia from Apr 4th 2011.[5] Technically speaking, a Talk page is a normal Wiki page located in one of the Talk namespaces. In this work, we focus on article Talk pages and do not regard User Talk pages. We selected the discussion pages according to the number of turns they contain. First, we discarded all discussion pages with less than four contributions. We then analyzed the distribution of turn counts per discussion page in the remaining set of pages and defined three classes: (i) discussion pages with 4-10 turns, (ii) pages with 11-20 turns, and (iii) pages with more than 20 turns. We then randomly extracted 50 discussion pages from class (i), 40 pages from class (ii) and 10 pages from class (iii). This decision is grounded in the restricted resources for the human annotation task.

Data Preprocessing Due to a lack of discussion structure, extracting the discussion threads from the Talk pages requires a substantial amount of preprocessing. Laniado et al. (2011) tackle the thread extraction by using text indentation and inserted user signatures as clues. We found these

[5]The snapshot contains 69900 articles and 5783 Talk pages of which 683 contained more than 3 contributions.

attributes to be insufficient for a reliable reconstruction of the thread structure.[6]

Our preprocessing approach consists of three steps: *data retrieval*, *topic segmentation* and *turn segmentation*. For retrieving the discussion pages, we use the *Java Wikipedia Library (JWPL)* (Zesch et al., 2008), which offers efficient, database-driven access to the contents of Wikipedia. We segment the individual Talk pages into discussions topics using the MediaWiki parser that comes with JWPL. In our corpus, the parser managed to identify all topic boundaries without any errors. The most complex preprocessing step is the turn segmentation.

First, we use the revision history of the Talk page to identify the author and the creation time of each paragraph. We use the *Wikipedia Revision Toolkit* (Ferschke et al., 2011) to examine the changes between adjacent revisions of the Talk page in order to identify the exact time a piece of text was added as well as the author of the contribution. We have to filter out malicious edits from the history, as they would negatively affect the segmentation process. We therefore disregard all edits that are reverted in later later revisions. In contrast to vandalism on article pages, this approach has proven to be sufficient to detect vandalism in the Talk page history.

Within each discussion topic, we aggregate all adjacent paragraphs with the same author and the same time stamp to one turn. In order to account for turns that were written in multiple revisions, we regard all time stamps within a window of 10 minutes[7] as belonging to the same turn, unless the page was edited by another user in the meantime. Finally, the turn is marked with the indentation level of its least indented paragraph. This information is used to identify the relationship between the turns, since indentation is used to indicate a reply to an existing comment in the discussion.

A co-author of this paper evaluated the acceptability of the boundaries of each turn in the SEWD corpus and found that 94% of the 1450 turns were correctly segmented. Turns with segmentation errors were not included in the gold standard.

[6] Viégas et al. (2007) reported that only 67% of the contributions on Wikipedia Talk pages are signed, which makes signatures an unreliable predictor for turn boundaries.

[7] We experimentally tested values between 1 and 60 minutes.

Annotation Process For our annotation study, we used the freely available MMAX2 annotation tool[8]. Two annotators were introduced to the annotation schema by an instructor and trained on an extra set of ten discussion pages. During the annotation of the corpus, the annotators were allowed to discuss difficult cases and could consult the instructor if in doubt. They had access to the segmented discussion pages within the MMAX2 tool as well as to the original Wikipedia articles and discussion pages on the web.

The reconciliation of the annotations was carried out by an expert annotator. In order to obtain a consolidated gold standard, the expert decided all cases in which the annotations of the two annotators did not match. Descriptive statistics for the label assignments of each annotator and for the gold standard can be seen in Table 2 and will be further discussed in Section 4.2.

Corpus Format We publish our SEWD corpus in two formats[9], the original MMAX format, and as XMI files for further processing with the *Apache Unstructured Information Management Architecture*[10]. For the latter format, we also provide the type system which defines all necessary corpus specific types needed for using the data in an NLP pipeline.

4.1 Inter-Annotator Agreement

To evaluate the reliability of our dataset, we perform a detailed inter-rater agreement study. For measuring the agreement of the individual labels, we report the observed agreement, Kappa statistics (Carletta, 1996), and F_1-scores. The latter are computed by treating one annotator as the gold standard and the other one as predictions (Hripcsak and Rothschild, 2005). The scores can be seen in Table 2.

The average observed agreement across all labels is $\bar{P_O} = .94$. The individual Kappa scores largely fall into the range that Landis and Koch (1977) regard as *substantial agreement*, while three labels are above the more strict .8 threshold for reliable annotations (Artstein and Poesio, 2008). Furthermore, we obtain an overall pooled Kappa (De Vries et al., 2008) of $\kappa_{pool} = .67$,

[8] http://www.mmax2.net

[9] http://www.ukp.tu-darmstadt.de/data/wikidiscourse

[10] http://uima.apache.org

Label	Annotator 1		Annotator 2		Inter-Annotator Agreement				Gold Standard	
	N	Percent	N	Percent	$N_{A_1 \cup A_2}$	P_O	κ	F_1	N	Percent
Article Criticism										
CM	183	13.4%	105	7.7%	193	.93	.63	.66	116	8.5%
CW	106	7.8%	57	4.2%	120	.95	.52	.55	70	5.1%
CU	69	5.0%	35	2.6%	83	.95	.38	.40	42	3.1%
CS	164	12.0%	101	7.4%	174	.94	.66	.69	136	9.9%
CL	195	14.3%	199	14.6%	244	.93	.73	.77	219	16.0%
COBJ	27	2.0%	23	1.7%	29	.99	.84	.84	27	2.0%
CO	20	1.5%	59	4.3%	71	.95	.18	.20	48	3.5%
Explicit Performative										
PSR	458	33.5%	351	25.7%	503	.86	.66	.76	406	29.7%
PREF	43	3.1%	31	2.3%	51	.98	.61	.62	45	3.3%
PFC	73	5.3%	65	4.8%	86	.98	.76	.77	77	5.6%
PPC	357	26.1%	340	24.9%	371	.97	.92	.94	358	26.2%
Information Content										
IP	1084	79.3%	1027	75.1%	1135	.89	.69	.93	1070	78.3%
IS	228	16.7%	208	15.2%	256	.95	.80	.83	220	16.1%
IC	187	13.7%	109	8.0%	221	.89	.46	.51	130	9.5%
Interpersonal										
ATT+	71	5.2%	140	10.2%	151	.94	.55	.58	144	10.5%
ATTP	71	5.2%	30	2.2%	79	.96	.42	.44	33	2.4%
ATT-	67	4.9%	74	5.4%	100	.96	.56	.58	87	6.4%

Table 2: Label frequencies and inter-annotator agreement. $N_{A_1 \cup A_2}$ denotes the number of turns that have been labeled with the given label by at least one annotator. P_O denotes the observed agreement.

which is defined as

$$\kappa_{pool} = \frac{\bar{P_O} - \bar{P_E}}{1 - \bar{P_E}} \quad (1)$$

with

$$\bar{P_O} = \frac{1}{L} \sum_{l=1}^{L} P_{O_l} \quad , \quad \bar{P_E} = \frac{1}{L} \sum_{l=1}^{L} P_{E_l} \quad (2)$$

where L denotes the number of labels, P_{E_l} the expected agreement and P_{O_l} the observed agreement of the l^{th} label. κ_{pool} is regarded to be more accurate than an averaged Kappa.

For assessing the overall inter-rater reliability of the label set assignments *per turn*, we chose Krippendorff's Alpha (Krippendorff, 1980) using MASI, a measure of agreement on set-valued items, as the distance function (Passonneau, 2006). MASI accounts for partial agreement if the label sets of both annotators overlap in at least one label. We achieved an Alpha score of $\alpha = .75$. According to Krippendorff, datasets with this score are considered reliable and allow tentative conclusions to be drawn.

The CO label showed the lowest agreement of only $\kappa = .18$. The label was supposed to cover any criticism that is not covered by a dedicated label. However, the annotators reported that they chose this label when they were unsure whether a particular criticism label would fit a certain turn or not.

Labels in the interpersonal category all show agreement scores below 0.6. It turned out that the annotators had a different understanding of these labels. While one annotator assigned the labels for any kind of positive or negative sentiment, the other used the labels to express agreement and disagreement between the participants of a discussion.

A common problem for all labels were contributions with a high degree of indirectness and implicitness. Indirect contributions have to be interpreted in the light of conversational implicature theory (Grice, 1975), which requires contextual knowledge for decoding the intentions of a speaker. For example, the message

Is population density allowed to be n/a?

has the surface form of a question. However, the context of the discussion revealed that the author tried to draw attention to the missing figure in the article and requested it to be filled or removed. The annotators rarely made use of the context, which was a major source for disagreement in the study.

Another difficulty for the annotators were long discussion turns. While the average turn consists of 42 tokens, the largest contribution in the corpus is 658 tokens long. Turns of this size can cover multiple aspects and potentially comprise many different dialog acts, which increases the probability of disagreement. This issue can be addressed by going from the turn level to the utterance level in future work.

A comparison of our results with the agreement reported for other datasets shows that the reliability of our annotations lies well within the field of the related work. Bender et al. (2011) carried out an annotation study of social acts in 365 discussions from 47 Wikipedia Talk pages. They report Kappa scores for thirteen labels in two categories ranging from .13 to .66 per label. The overall agreement for each category was .50 and .59, respectively, which is considerably lower than our $\kappa_{pool} = .67$. Kim et al. (2010b) annotate pairs of posts taken from an online forum. They use a dialog act tagset with twelve labels customized for modeling troubleshooting-oriented forum discussions. For their corpus of 1334 posts, they report an overall Kappa of .59. Kim et al. (2010a) identify unresolved discussions in student online forums by annotating 1135 posts with five different speech acts. They report Kappa scores per speech act between .72 and .94. Their better results might be due to a more coarse grained label set.

4.2 Corpus Analysis

The SEWD corpus contains 313 discussions consisting of 1367 turns by 337 users. The average length of a turn is 42 words. 208 of the 337 contributors are registered Wikipedia users, 129 wrote anonymously. On average, each contributor wrote 168 words in 4 turns. However, there was a cluster of 16 people with ≥ 20 contributions.

Table 2 shows the frequencies of all labels in the SEWD corpus. The most frequent labels are *information providing* (IP), *requests* (PSR) and *reports of performed edits* (PPC). The IP-label was assigned to more than 78% of all 1367 turns, because almost every contribution provides a certain amount of information. The label was only omitted if a turn merely consisted of a discussion template but did not contain any text or if it exclusively contained questions.

More than a quarter of the turns are labeled with PSR and PPC, respectively. This indicates that edit requests and reports of performed edits are the main subject of discussion. Generally, it is more common that edits are reported after they have been made than to announce them before they are carried out, as can be seen in the ratio of PPC to PFC labels. The number of turns labeled with PSR is almost the same as the number of contributions labeled with either PPC or PFC. This allows the tentative conclusion that nearly all requests potentially lead to an edit action. As a matter of fact, the most common label adjacency pair[11] in the corpus is PSR→PPC, which substantiates this assumption.

Article criticism labels have been assigned to 39.4% of all turns. Almost half (241) of the labels from this class are assigned to the first turn of a discussion. This shows that it is common to open a discussion in reference to a particular deficiency of the article. The large number of CL labels compared to other labels from the same category is due to the fact that the Simple English Wikipedia requires authors to write articles in a way that they are understandable for non-native speakers of English. Therefore, the use of adequate language is one of the major concerns of the Simple English Wikipedia community.

5 Automatic Dialog Act Classification

For the automatic classification of dialog acts in Wikipedia Talk pages, we transform the multi-label classification problem into a binary classification task (Tsoumakas et al., 2010). We train a binary classifier for each label using the WEKA data-mining software (Hall et al., 2009). We use three learners for the classification task, a Naive Bayes classifier, J48, an implementation of the C4.5 decision tree algorithm (Quinlan, 1992) and SMO, an optimization algorithm for training support vector machines (Platt, 1998). Finally, we combine the best performing learners for each label in a UIMA-based classification pipeline (Ferrucci and Lally, 2004).

Features for Dialog Act Classification As features, we use all uni-, bi- and trigrams that occurred in at least three different turns. Furthermore, we include the time distance to the previous and the next turn (in seconds), the length of the current, previous and next turn (in tokens), the

[11] A label transition $A \rightarrow B$ is recorded if two adjacent turns are labeled with A and B, respectively.

position of the turn within the discussion, the indentation level of the turn and two binary features indicating whether a turn references or is referenced by another turn.[12] In order to capture the sequential nature of the discussions, we use the n-grams of the previous and the next turn as additional features.

Balancing Positive and Negative Instances Since the number of positive instances for each label is small compared to the number of negative instances, we create a balanced dataset which contains an equal amount of positive and negative instances. Therefore, we randomly select the appropriate number of negative instances and discard the rest. This improves the classification performance on every label for all three learners.

Feature Selection Using the full set of features, we achieve the following macro/micro averaged F_1-scores: 0.29 / 0.57 for Naive Bayes, 0.42 / 0.66 for J48 and 0.43 / 0.72 for SMO. To further improve the classification performance, we reduce the feature space using two feature selection techniques, the χ^2 metric (Yang and Pedersen, 1997) and the Information Gain approach (Mitchell, 1997). For each label, we train separate classifiers using the top 100, 200 and 300 features obtained by each feature selection technique and choose the best performing set for our final classification pipeline.

Indentation and *temporal distance to the preceding turn* proved to be the best ranked non-lexical features overall. Additionally, the *turn position within the topic* was a crucial feature for most labels in the criticism class and for `PSR` and `IS` labels. This is not surprising, because article criticism, suggestions and questions tend to occur in the beginning of a discussion. The two *reference* features have not proven to be useful. The relational information was better covered by the *indentation* feature. The subjective quality of the lexical features seems to be correlated with the inter-annotator agreement of the respective labels. Features for labels with low agreement contain many n-grams without any recognizable semantic connection to the label. For labels with good agreement, the feature lists almost exclusively contain meaningful lexical cues.

[12] A turn Y references a preceding turn X if the indentation level of Y is one level deeper than of X.

Label	Human	Base	Naive Bayes	J48	SMO	Best
CM	.66	.07	.68	.48	.66	.68
CW	.55	.01	.70	.20	.56	.70
CU	.40	.07	.66	.35	.59	.66
CS	.69	.09	.67	.67	.75	.75
CL	.77	.11	.70	.66	.73	.73
COBJ	.84	.04	.78	.51	.63	.78
CO	.20	.02	.61	.06	.39	.61
PSR	.76	.30	.72	.70	.76	.76
PREF	.62	.00	.76	.41	.64	.76
PFC	.77	.04	.70	.62	.73	.73
PPC	.94	.25	.74	.82	.85	.85
IP	.93	.74	.83	.93	.93	.93
IS	.83	.16	.79	.86	.85	.86
IC	.51	.06	.67	.32	.59	.67
ATT+	.58	.10	.61	.65	.72	.72
ATTP	.44	.03	.72	.25	.62	.72
ATT-	.58	.07	.52	.30	.52	.52
Macro	.65	.13	.70	.52	.68	.73
Micro	.79	.35	.74	.75	.80	.82

Table 3: F_1-Scores for the balanced set with feature selection on 10-fold cross-validation. *Base* refers to the baseline performance, *Best* to our classification pipeline.

Classification Results Table 3 shows the performance of all classifiers and our final classification pipeline evaluated on 10-fold cross validation. Naive Bayes performed surprisingly well and showed the best macro averaged scores among the three learners while SMO showed the best micro averaged performance. We compare our results to a random baseline and to the performance of the human annotators (cf. Table 3 and Figure 2). The baseline assigns the dialog act labels at random according to their frequency distribution in the gold standard. Our classifier outperformed the baseline significantly on all labels.

The comparison with the human performance shows that our system is able to reach the human performance. In most cases, the annotation agreement is reliable, and so are the results of the automatic classification. For the labels `CU` and `CO`, the inter-annotator agreement is not high. The comparably good performance of the classifiers on these labels shows that the instances do have shared characteristics. Human raters, however, have difficulties recognizing these labels consistently. Thus, their definitions need to be refined in future work.

To our knowledge, none of the related work on discourse analysis of Wikipedia Talk pages per-

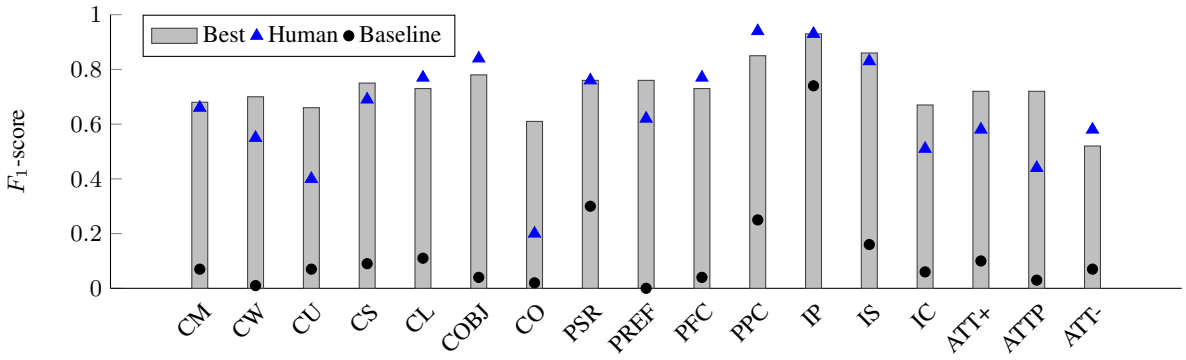

Figure 2: F_1-Scores for our classification pipeline (*Best*), the human performance and baseline performance.

formed automatic dialog act classification. However, there has been previous work on classifying speech acts in other discourse types. Kim et al. (2010a) use Support Vector Machines (SVM) and Transformation Based Learning (TBL) for the automatic assignment of five speech acts to posts taken from student online forums. They report individual F_1-scores per label which result in a macro average of 0.59 for SVM and 0.66 for TBL. Cohen et al. (2004) classify speech acts in emails. They train five binary classifiers using several learners on 1375 emails and report F_1 scores per speech act between .44 and .85. Despite the larger tagset, our classification approach achieves an average F_1-score of .82 and therefore lies in the top ranks of the related work.

6 Conclusions

In this paper, we proposed an annotation schema for the discourse analysis of Wikipedia discussions aimed at the coordination efforts for article improvement. We applied the annotation schema to a corpus of 100 Wikipedia Talk pages, which we make freely available for download. A thorough analysis of the inter-annotator agreement showed that the dataset is reliable. Finally, we performed automatic dialog act classification on Wikipedia Talk pages. Therefore, we combined three machine learning algorithms and two feature selection techniques to a classification pipeline, which we trained on our SEWD corpus. We achieve an average F_1-score of .82, which is comparable to the human performance of .79. The ability to automatically classify discussion pages will help to investigate the relations between article discussions and article edits, which is an important step towards understanding the processes of collaboration in large-scale Wikis. Furthermore, it will be the basis for practical applications that bring the hidden content of Talk pages to the attention of article readers.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the Hessian research excellence program "Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" (LOEWE) as part of the research center "Digital Humanities".

References

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, December.

John L. Austin. 1962. *How to Do Things with Words*. Clarendon Press, Cambridge, UK.

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating Social Acts: Authority Claims and Alignment Moves in Wikipedia Talk Pages. In *Proceedings of the Workshop on Language in Social Media*, pages 48–57, Portland, Oregon, USA.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.

Tamitha Carpenter and Emi Fujioka. 2011. The Role and Identification of Dialog Acts in Online Chat. In *Proceesings of the Workshop on Analyzing Microtext at the 25th AAAI Conference on Artificial Intelligence*, San Francisco, CA, USA.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to Classify Email into "Speech Acts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 309–316, Barcelona, ES.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *Proceedings of the Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Cambridge, MA, USA.

Han De Vries, Marc N. Elliott, David E. Kanouse, and Stephanie S. Teleki. 2008. Using Pooled Kappa to Summarize Interrater Agreement across Many Items. *Field Methods*, 20(3):272–282.

David Ferrucci and Adam Lally. 2004. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10:327–348.

Oliver Ferschke, Torsten Zesch, and Iryna Gurevych. 2011. Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia's Edit History. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations*, pages 97–102, Portland, OR, USA.

Paul Grice. 1975. Logic and Conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics*, volume 3. New York: Academic Press.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11:10–18.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Dan Jurafsky, Liz Shriberg, and Debbra Biasca. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. Technical Report Draft 13, University of Colorado, Institute of Cognitive Science.

Jihie Kim, Jia Li, and Taehwan Kim. 2010a. Towards Identifying Unresolved Discussions in Student Online Forums. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 84–91, Los Angeles, CA, USA.

Su Nam Kim, Li Wang, and Timothy Baldwin. 2010b. Tagging and linking web forum posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 192–202, Stroudsburg, PA, USA.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage Publications.

J. Richard Landis and Gary G. Koch. 1977. An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics*, 33(2):363–374, June.

David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. 2011. When the Wikipedians Talk: Network and Tree Structure of Wikipedia Discussion Pages. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, Dublin, IE.

Tom Mitchell. 1997. *Machine Learning*. McGraw-Hill Education (ISE Editions), 1st edition.

Rebecca Passonneau. 2006. Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, IT.

John C. Platt. 1998. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*, pages 185–208, Cambridge, MA, USA.

Ilona R. Posner and Ronald M. Baecker. 1992. How People Write Together. In *Proceedings of the 25th Hawaii International Conference on System Sciences*, pages 127–138, Wailea, Maui, HI, USA.

Ross Quinlan. 1992. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1st edition.

Jodi Schneider, Alexandre Passant, and John G. Breslin. 2011. Understanding and Improving Wikipedia Article Discussion Spaces. In *Proceedings of the 26th Symposium on Applied Computing*, Taichung, TW.

John R. Searle. 1969. *Speech Acts*. Cambridge University Press, Cambridge, UK.

John R. Searle. 1976. A classification of illocutionary acts. *Language in Society*, 5:1–23.

Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. 2008. Information Quality Work Organization in Wikipedia. *Journal of the American Society for Information Science*, 59:983–1001.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis P. Vlahavas. 2010. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer.

Fernanda Viégas, Martin Wattenberg, Jesse Kriss, and Frank Ham. 2007. Talk Before You Type: Coordination in Wikipedia. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, Waikoloa, Big Island, HI, USA.

Eti Yaari, Shifra Baruchson-Arbib, and Judit Bar-Ilan. 2011. Information quality assessment of community generated content: A user study of Wikipedia. *Journal of Information Science*, 37:487–498.

Yiming Yang and Jan O. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420, San Francisco, CA, USA.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, MA.